IBM'S LVCSR SYSTEM FOR TRANSCRIPTION OF BROADCAST NEWS USED IN THE 1997 HUB4 ENGLISH EVALUATION

S. Chen, M. J. F. Gales, P. S. Gopalakrishnan, R. A. Gopinath, H. Printz, D. Kanevsky, P. Olsen and L. Polymenakos

IBM T. J. Watson Research Center, Yorktown Heights, NY.,
email:rameshg@watson.ibm.com, phone: (914)-945-2794

ABSTRACT

This paper describes IBM's large vocabulary continuous speech recognition (LVCSR) system used in the 1997 Hub4 English evaluation. It focusses on extensions and improvements to the system used in the 1996 evaluation. The recognizer uses an additional 35 hours of training data over the one used in the 1996 Hub4 evaluation [8]. It includes a number of new features: optimal feature space for acoustic modeling (in training and/or testing), filler-word modeling, Bayesian Information Criterion (BIC) based segmentation and segment clustering, an improved implementation of iterative MLLR, variance adaptation, and 4-gram language models. Results using the 1996 and 1997 DARPA Hub4 evaluation data sets are presented.

1. INTRODUCTION

Recently interest in large vocabulary continuous speech recognition (LVCSR) research has shifted from read speech data to speech data found in the real world - like broadcast news (BN) over radio and TV and conversational speech over the telephone. Considerable amount of both acoustic (approximately 100 hours of which about 70% is usable) and linguistic (approximately 400 million words) training data for BN has been made by the Linguistic Data Consortium (LDC) in the context of DARPA sponsored Hub4 evaluations of LVCSR systems on BN [1]. As has been studied and reported by several researchers [4, 8, 12, 11, 9, 10], BN transcription poses several challenges to LVCSR systems. The speech data exhibits a wide variety of speaking styles, environmental and background noise conditions and channel conditions. A popular approach has been to classify the BN data into a set of homogeneous conditions and to build acoustic models (AMs) for each condition. Test data is then segmented and classified along conditions and an appropriate acoustic model used for each condition. One particular classification scheme for BN news data that has been used in the DARPA sponsored Hub4 BN evaluation in 1996 splits the speech data along the so-called F-conditions [1]: prepared speech (F0), spontaneous speech (F1), low fidelity speech, including telephone channel speech (F2), speech in the presence of background music (F3), speech in the presence of background noise (F4), speech from non-native speakers (F5) and FX - all other speech. The 1996 Hub4 Unpartitioned Evaluation (UE) and Partitioned Evaluation (PE) test data set forms a standard test set for evaluating LVCSR systems. The only difference between the UE and PE tests is that in the latter, the data is segmented and classified into F-conditions manually, while in the former this has to be done automatically if necessary. A comparison gives information on how well automatic segmentation schemes work for BN news transcription.

For the UE test, in the past we have used a two-stage approach [4]. The speech data is first segmented into high bandwidth speech (clean), low bandwidth speech (telephone), and music. The music segments are removed and the high and low bandwidth speech segments are then decoded using models trained (or adapted) on high and low bandwidth speech respectively. This is because of the inadequacy of current segmentation algorithms to separate out other F-conditions; it is relatively easy to detect music or telephone channel speech.

For the PE test, in the past we have built condition specific models for each condition using MAP and MLLR. This is because there is not sufficient training data to independently build models for each F-condition; besides, it may not be the best way to handle the problem.

Our current approach for both the UE and PE tests is to use a single robust model built on all the available training data. Speaker/condition-adapted (SAT) training [6], while appropriate for this purpose, is not used in the model described in this paper. For both the PE and UE tests, iterative MLLR is used to adapt the baseline robust model for both the speaker and the F-condition. For the UE test, the data is still, however, segmented into low-bandwidth and high-bandwidth segments. The segments are then clustered into homogeneous groups (the same speaker or environmental condition) before iterative MLLR is applied.

The 1997 DARPA Hub4 evaluation test was a UE test. However, after the evaluation the distributed reference scripts was used to also form a PE test. This paper reports numbers on both these tests.

In the following sections we present algorithmic improvements to the baseline model that were used in the 1997 Hub4 UE evaluation and the post-evaluation PE test (setup as described above). The focus of our research effort has been on improving baseline recognition accuracy for clean speech (i.e., the F0 and F1 conditions). Nearly all of the algorithm development work was evaluated only on these two conditions.

2. OVERVIEW OF THE LVCSR SYSTEM

The IBM LVCSR system uses acoustic models for subphonetic units with context-dependent tying (see [2, 3] for details). Context dependent sub-phone class instances are identified by growing a decision tree from the available training data [2] and specifying the terminal nodes of the tree as the relevant instances of these classes. The acoustic feature vectors that characterize the training data at the leaves are modeled by a mixture of Gaussian pdf's, with diagonal covariance matrices. The HMM used to model each leaf is a simple 1-state model, with a self-loop and a forward transition.

The recognizer used in the 1996 evaluation had 5.7K HMM states (or leaves) and 170K gaussians. The decision tree for the HMM states was built WSJ0+1 data. The gaussian mixtures, however were trained on the approximately 35 hours of BN training data distributed by LDC

Acoustic Model	F0	F1
AM-base	21.4	30.3
AM-0(4.0K)	21.3	29.7
AM-1(2.0K)	22.6	31.0
AM-2(3.5K)	21.1	29.1
AM-3(7.3K)	21.9	30.3

Table 1. Comparison of Decision Tree Sizes: AM-base - trained on WSJ, AM-0 - trained on BN with 4K leaves, AM-1 trained on F0+F1 portion with 2K leaves, AM-2 same as AM-1 with 3.5K leaves, AM-3 same as AM-2 with 7.3K leaves.

in 1996. For the PE test, the models were further adapted to each focus condition and for the UE test to high/low bandwidth speech using a combination of MAP and MLLR [17, 16, 8] adaptation.

The recognizer used in this year's (1997) evaluation had 3.5K HMM states and 170K gaussians. The decision trees were built exclusively on the F0 and F1 portions of the complete 70 hours of BN training data. There was a single robust acoustic model (AM-Evl97: built on all the data and adapted in a supervised fashion to F0 and F1 data using MLLR) used for baseline decoding which was followed by iterative MLLR.

3. ACOUSTIC MODELING

We began our effort by first building a new baseline acoustic model (AM-base) with 90K gaussians (the smaller size was preferred to run experiments quickly) using all the 70 hours of training data (including the 35 hours of additional data distributed in 1997) by rebuilding gaussian mixtures for the 5741 HMM states of our 1996 evaluation acoustic models. Initial experiments indicated that there was very little ($< .5\%$) in WER by using the extra 35 hours of data. Since these HMM states in these models were constructed from WSJ data we built two new decision trees for context clustering, one based on just the clean (F0+F1) training data and the other based on all the training data. Gaussian mixtures were then estimated using the EM algorithm and the performance for various model sizes were evaluated. Experimental results for the F0 and F1 focus conditions on the 1996 PE test are shown in Table 1. The language model (LM) used in these experiments is LM-base (see below) and there are about 90K gaussians in each of the acoustic models. Firstly, notice that building the decision tree with the BN data improves error rate on both F0 and F1 (WER with AM-base is worse than WER with AM-0 or AM-2). The improvements are more on F1 (spontaneous speech) because of the new realizations of context-dependent subphonetic units vis à vis WSJ training data. Secondly, not using the training data for the other F-conditions in tree building gives more gain (AM-0 vs. AM-2). This is probably because some of the HMM states are now modeling realizations of phones in specific environmental conditions. The best results were obtained with a system with about 3.5K HMM states (AM-2). These were the HMM states used in AM-Evl97, the acoustic models used in the evaluation.

3.1. Filler Models

The training data has been transcribed with breath and filled-pauses. This allows us to build models for filler words. Filler words are transcribed using our usual phone set of 51 phones in the dictionary. To the decision trees that take sub-phonetic units to the HMM states, new states were added for each occurrence of a phone within a particular filler word. The models for these states were initialized by those of randomly chosen state from the same sub-phonetic unit. Standard Baum-Welch reestimation is then used to estimate the models. Filler models seemed to improve the performance on spontaneous speech without degrading the performance on prepared speech when the base models was AM-base. However, the gain was marginal when the base model was AM-2. This is presumably because the MM states in AM-base were built on WSJ data while the HMM states in AM-2 were built on the BN training data (that had 50% of F1 data - where filled paused usually occur) and hence some states were presumably already modeling filler words. Results are summarized in Table 2. The HMM states in AM-Evl97 are the same as the ones in AM-2; filler models were *not* used in the 1997 Hub4 evaluation.

Acoustic Model	F0	F1
AM-base	21.4	30.3
AM-4	21.0	29.0
AM-2	21.1	29.1
AM-5	21.0	28.9

Table 2. % word error rate with filler word models: AM-base and AM-2 do not use filler models. AM-4 is AM-base with filler models and AM-5 is AM-2 with filler models.

3.2. Optimal Features Spaces for Modeling

The next acoustic model improvement came from finding optimal features for modeling. The motivation is the following: the number of gaussians used in current LVCSR systems forces us (from data insufficiency, storage and computational considerations) to typically use diagonal gaussian models. Meanwhile, it is clear that with full-covariance gaussian models, linear transformations of the feature space do not lead to better models. Moreover, if the transformation is unimodular (or volume-preserving) the likelihood is exactly the same in all transformed spaces. However, with diagonal gaussian models one can ask the following question: among all possible transformed feature spaces which is the one where the diagonal assumption is "most valid". What do we mean by "most valid"? If the transformation is unimodular (required only to simplify the argument), then, in each transformed space there is a loss in likelihood with respect to full-covariance modeling (which is a constant). One can therefore find a transformed space in which the loss in likelihood is least (for details see [13, 22, 20, 14]). One of us, R. Gopinath, was exposed to this idea based on N. Kumar's Phd Thesis [18]. The viewpoint in that paper was to find a generalization of LDA by allowing a more realistic assumption on the covariances. Ignoring projections, his work directly translates to the "best feature space for diagonal modeling" described here. He suggests using a numerical scheme for the obtaining the transformation. The same idea was independently developed by M. Gales (one of the authors!), while he was at Cambridge University under the name "semi-tied covariances" [20, 14]. The viewpoint there is to have covariances of the form AD_jA^T (with D_j diagonal) for each of the gaussians motivated by the fact that correlations can be better modeled this way. The matrix A is typically shared by a collection of gaussians. Gales also gives efficient numerical algorithms to compute A. Semi-tied covariances and finding the best class-dependent feature space for modeling with diagonal gaussians are essentially the same idea.

In summary, to better model the data with diagonal gaussians, one can use a single global transformation of the feature space. Notice however, that the gaussians can be clustered into groups and each group can be modeled in its own feature space. Since there is more flexibility in this case the loss in likelihood is less. In the extreme case where each gaussian has its own feature space transformation one can

Acoustic Model	F0	F1
AM-2(baseline)	21.1	29.1
AM-6(1 transform)	19.3	28.4
AM-7(4 transforms)	19.4	29.0

Table 3. Optimal Feature Spaces for HMM state clusters: a) AM-2 - baseline b) AM-6 - single transform c) AM-7 - 4 transforms

choose the transformation to be projection onto the eigenbasis of its covariance matrix and the likelihood of the data is the same as full-covariance likelihood. However, from computational and storage points of view this is exactly as expensive as full-covariance modeling.

The optimal feature space idea was tried on our Hub4 recognizer which had 3.5K HMM states. For the purposes of finding the transformation, each state was modeled by a single gaussian in $\mathbb{R}^{60}$ obtained by double-rotation (a variant of LDA) of cepstral features derived from the speech data [7]. The training data consisted of $N \approx 24M$ labeled samples. If (x_i, l_i) is the labeled (at HMM state level) training data, $i \in \{1, 2, \ldots, N\}, x_i \in \mathbb{R}^d, l_i \in \{1, 2, \ldots, J\}$, and $c_j \in \{1, 2, \ldots, K\}$ is the class cluster (or transformation id) map, and Σ_j is the covariance at state j (we are assuming a single gaussian at each state for simplicity), then the likelihood of the training data with the single gaussian assumption is given by the following expression [13]:

$$p^{\star}_{diag}(x_1^N) = g(N, d) \prod_{j=1}^{J} \left| A_{c_j} \right|^{N_j} \left| diag(A_{c_j} \bar{\Sigma}_j A_{c_j}^T) \right|^{-\frac{N_j}{2}}.$$

Maximizing the above expression numerically gives the optimal choice of transforms $A_k, k \in \{1, 2, \ldots, K\}$. We experimented with class clusters obtained by data-driven clustering of HMM states and by knowledge-based sharing (e.g., all HMM states of a phone share the same A_k).

In our experiments, after the transform is obtained this way, using single-pass-retraining from a baseline system, gaussian mixture models are built for each HMM state using the new (state-dependent) feature space. Here we present results when using one (AM-6) and four transformations (AM-7) on the AM-2 baseline acoustic models (see Table 3). For the latter case one transform each was used for all the gaussians corresponding to a) stop-consonants and flaps, b) fricatives, c) vowels and dipthongs, and d) nasals, glides, and silence respectively. Based on these results, and also because iterative MLLR is greatly simplified if there is only a single transformation, AM-Evl97 used the feature space transformation in AM-6.

3.3. More Mixture Components

The AM-6 models were further enhanced by increasing the number of components to give about 170K gaussians using standard clustering and EM resestimation in the optimal feature space. The results using these models (AM-8) are shown in Table 4. It turns out that combined scoring gives better error rates than scoring each condition separately due to some artifacts of the scoring process. Therefore, the combined scoring numbers for AM-8 (AM-8 combined) are also given in Table 4.

3.4. Supervised adaptation F0 and F1

All of the acoustic models above were built on training data from all the F-conditions. Since we were especially interested in the performance of our LVCSR system on F0 and F1 the model AM-8 was further adapted using the F0 and F1 portion of the training data (about 60%) in a supervised fashion using MLLR. The performance of the baseline

Acoustic Model	F0	F1
AM-6	19.3	28.4
AM-8	17.7	26.2
AM-8 (combined)	17.5	24.9

Table 4. Increased number of mixture components for HMM states: a) AM-6 - 90K Gaussians b) AM-8 - 170K Gaussians

model (AM-8) and the adapted model (AM-9) are shown in Table 5 for all the F-conditions. Notice that the F0, F1 and FX error rates are better with supervised adaptation, while F2 and F3 and F4 degrades somewhat. These experiments used the LM used in the evaluation.

4. SEGMENTATION AND CLUSTERING

4.1. Segmentation

The segmentation algorithm evolved continually till the final evaluation. Consequently, the algorithm used in the evaluation has some components that remained for historical reasons. Initially the plan was to use Gaussian mixture models for low bandwidth speech, high bandwidth speech and pure music to segment the data as we had done before [4]. The reason for separating telephone bandwidth speech was to avoid mixing up high and low bandwidth speech in unsupervised adaptation. Besides, to ensure that the segment boundaries occur at silences, the test data is also decoded with a small model set (we used AM-6). The silence information from this decode pass is used to prevent segment boundaries from splitting words. data is identical (except for the side-information of the F-conditions in PE). Therefore a comparison of UE and PE performance gives and evaluation of the segmentation procedure. Experimental tests were conducted on the acoustic model AM-6 described earlier and the results are shown in Table 6 The segmentation procedure typically leads to an overall loss of about 1% absolute in WER. The segments from the above procedure typically contain data from both male and female speakers. This may be undesirable for both baseline decoding and unsupervised adaptation. Gaussian mixture gender models were built and added to the segmentation process above. Just days before the evaluation, a speaker change detection scheme was tried to see if speaker turns can be detected. This was again motivated by potential improvements to unsupervised adaptation. The scheme uses Bayesian Information Criterion (which is essentially a penalized ML scheme) to find penalized ML change points in the speech signal. This scheme works very well in detecting speaker, background, and channel changes. Therefore, in principle, this scheme subsumes the gender detection, channel detection and music detection schemes described earlier and moreover does not rely on building models of various conditions. However, because there was not sufficient time to study this scheme, it was used as a pre-processing step in the final evaluation segmentation. The evaluation speech data was split into "turns" using this turn detector and each turn was then subsequently segmented further using telephone/clean/music detectors and the gender detector. The performance of this scheme on the 1996 evaluation data using the final baseline model AM-Evl97 is shown in Table 7. Using the reference scripts for the 1997 evaluation which is marked with speaker turns, a PE test was setup. This gives us a comparison of the segmentation accuracy on the 1997 evaluation data (see Table 8). There was a bug in the silence decode (part of the segmentation process) used in the actual evaluation fixing which improves the baseline UE by .5%.

Acoustic Model	Total	F0	F1	F2	F3	F4	F5	FX
AM-8	27.7	17.5	24.9	34.6	23.8	35.1	25.4	53.5
AM-Evl97	27.4	17.3	24.2	35.2	24.3	35.3	26.1	52.7

Table 5. Supervised adaptation on 1996 Evaluation data using MLLR (% WER): AM-8 - baseline, AM-Evl97 - adapted models.

Test	Total	F0	F1	F2	F3	F4	F5	FX
PE	28.2	18.6	25.1	34.8	24.7	34.8	29.1	54.2
UE	29.5	19.4	26.0	39.0	27.2	36.2	24.1	55.2

Table 6. Segmentation Accuracy on 1996 Evaluation data with AM-6 models: PE vs. UE (% WER).

Test	Total	F0	F1	F2	F3	F4	F5	FX
PE	27.4	17.3	24.2	35.2	24.3	35.3	26.1	52.7
UE	28.7	18.5	25.3	39.5	25.8	35.6	26.4	53.7

Table 7. Segmentation Accuracy on 1996 Evaluation data with AM-Evl97 models: PE vs. UE (% WER).

Test	Total	F0	F1	F2	F3	F4	F5	FX
PE	19.8	12.1	19.6	29.9	25.8	25.9	22.3	38.7
UE (Eval)	20.9	13.0	20.0	31.7	27.2	26.4	20.7	43.1
UE (Fixed)	20.4	12.4	19.9	30.5	26.9	26.0	21.8	43.5

Table 8. Segmentation Accuracy on 1997 Evaluation data with AM-Evl97 models: PE vs. UE (% WER).

Test	Total	F0	F1	F2	F3	F4	F5	FX
Baseline	20.9	13.0	20.0	31.7	27.2	26.4	20.7	43.1
MLLR+var	18.8	11.7	18.4	26.9	24.6	23.0	19.7	41.1
Iter. MLLR	18.0	11.2	17.9	25.3	24.3	22.4	19.1	39.3

Table 9. Unsupervised Adaptation on 1997 Evaluation data with AM-Evl97 models: Baseline vs. MLLR with efficient full-variance transformation vs. Iterative MLLR).

Test	Total	F0	F1	F2	F3	F4	F5	FX
Baseline	28.7	18.5	25.3	39.5	25.8	35.6	26.4	53.7
MLLR+var	26.6	17.7	24.3	33.0	21.8	34.1	25.8	48.5
Iter. MLLR	26.2	16.5	24.2	32.5	23.9	34.1	23.7	47.7

Table 10. Unsupervised Adaptation on 1996 Evaluation data with AM-Evl97 models: Baseline vs. MLLR with efficient full-variance transformation vs. Iterative MLLR).

Acoustic Model	F0	F1
AM-6+true cluster	17.5	24.8
AM-6+auto cluster	17.5	24.6

Table 12. Manual vs Automatic Clustering Performance

4.2. Unsupervised Adaptation on Test Data

Adaptation schemes like MLLR [16] adapt the means and variances of the gaussian models using linear transformations. If there are too many adaptation parameters or too little adaptation data, then, the adaptation tends to learn the adaptation data transcriptions quickly. To alleviate this problem we can decrease the number of adaptation parameters or increase the amount of adaptation data. The former is accomplished in the context of an iterative MLLR scheme where there are $2^i + 1$ transforms at the i^{th} iteration for 2^i non-silence phonetic sub-units and one transformation all the phonetic sub-units of silence. In the zeroth iteration an efficient full-variance linear transformation is estimated [19]. The basic idea there is to find a matrix A in an ML fashion on the test data such that the covariances are of the form AD_jA^T (D_j is the diagonal covariance obtained from training). Subsequently, in each iteration only the means are transformed. Increase in the amount of adaptation data is accomplished by clustering together similar the segments using a Bayesian Information Criterion (BIC) [15]. The results of unsupervised adaptation on the 1996 and 1997 Evaluation test sets are shown in Tables 10 and 9 respectively. In both cases the UE segments were clustered using the Bayesian Information Criterion as described below.

The 1997 evaluation system submission had a bug in the silence decode portion as described earlier. Unsupervised adaptation on the 1997 evaluation data after fixing this bug is shown in Table 11.

4.3. Clustering for Unsupervised Adaptation

The segments are clustered using a standard maximum-linkage bottom-up-clustering procedure with a single gaussian model for each segment and log-likelihood ratio distance measure. The termination for this bottom-up-clustering procedure was determined to maximize the BIC criterion [15, 21]. BIC is a likelihood criterion penalized by the model complexity (the number of parameters in the model). At each stage in the bottom-up-clustering process the increase in BIC value is computed and the process is terminated when this increase is negative. It can be easily be shown that the increase in BIC value by merging two clusters is given by

$$-n \log |\Sigma| + n_1 \log |\Sigma_1| + n_2 \log |\Sigma_2| + N(d + \frac{d(d+1)}{2}),$$

where $n = n_1 + n_2$ is sample size of the merged node, Σ is the covariance matrix of the merged node and N is the total number of samples from all the segments. This gives, in principle, a threshold-free approach to clustering.

To study the effectiveness of clustering, the F0 and F1 segments of PE test were clustered by hand (28 clusters) and by using the algorithm described above (31 clusters). The word error rate (WER) after iterative MLLR adaptation is nearly the same as seen in Table 12. In contrast the result of clustering all the PE segments automatically (79 clusters) is shown in Table 13 with single and multiple iterations of MLLR. For comparison the baseline numbers are also given.

5. LANGUAGE MODELING

The Language Model has a vocabulary of 65K most frequent words from the BN language model corpus distributed by LDC in 1996. The baseline language model (LM-base) is the one used in the 1996 evaluation [4]. With the same training data a standard 4-gram deleted interpolation LM was built (LM-4g). This component was added to LM-base to create LM-base+4g. This LM was further mixed with a small LM built from the 70 hours of acoustic training data transcriptions (LM-base+4g+ac). Mixing the 4-gram LM and the acoustic transcriptions LM to the baseline LM gives minor improvements to the recognition performance as seen in Table 14. The acoustic model used in these experiments was AM-6. LM-base+4g+ac was the language model used in the 1997 evaluation.

Lang. Model	F0	F1
LM-base	21.0	29.1
LM-base+4g	20.8	28.7
LM-base+4g+ac	20.7	28.6

Table 14. Mixture LM with 4-gram and acoustic transcriptions

6. CONCLUSION

Transcription of broadcast news poses several challenges. This paper presented IBM's LVCSR system used in the 1997 DARPA Hub4 evaluation. It systematically described all the changes that were incorporated into the 1996 evaluation system that lead to the system used the 1997 DARPA evaluation. The system was specifically built for handling baseline clean speech that is either read or conversational. A simple linear transformation into an optimal feature space for modeling is shown to lead to significant improvements in the baseline accuracy. New segmentation and clustering algorithms are used which significantly reduces the WER differential between partitioned and unpartitioned evaluations (currently about 1%). Unsupervised adaptation with a new clustering scheme gives about 10% relative improvement in accuracy. However, further improvements are required, especially, with regard to robustness to channel and noise degradations.

Acknowledgement This work was supported by DARPA in part under contract #DABT63-94-C0042 and in part under contract #MDA972-97-C-0012.

REFERENCES

[1] D. Pallet, "Overview of the 1997 DARPA Speech Recognition Workshop", Proc. of DARPA Speech Recognition Workshop, Feb 2-5, Chantilly VA, 1997.

[2] L. R. Bahl et al., "Robust Methods for using Context-Dependent features and models in a continuous speech recognizer", Proc. ICASSP, 1994.

[3] L. R. Bahl et al., "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task", Proc. ICASSP, pp 41-44, 1995.

[4] P. S. Gopalakrishnan et al., "Transcription of Radio Broadcast News with the IBM Large Vocabulary Speech Recognition System," Proc. ARPA SLT Workshop, Feb 1996.

[5] P. S. Gopalakrishnan, et al., "Acoustic Models Used in the IBM System for the ARPA Hub 4 Task," Proc. ARPA SLT Workshop, Feb 1996.

[6] T. Anastasakos, et al., "A Compact Model for Speaker-Adaptive Training", Proc. ICSLP-96.

[7] L. Bahl et. al., "Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA NAB News Task", Proc. SLT Workshop, Austin, TX, 1995.

Test	Total	F0	F1	F2	F3	F4	F5	FX
Baseline	20.4	12.4	19.9	30.5	26.9	26.0	21.8	43.5
MLLR+var	18.3	11.2	18.5	25.4	24.6	22.4	19.8	40.6
Iter. MLLR	17.7	10.8	17.6	23.9	25.4	21.5	19.7	41.9

Table 11. Unsupervised Adaptation on 1997 Evaluation data with AM-Evl97 models with segmentation bug fix: Baseline vs. MLLR with efficient full-variance transformation vs. Iterative MLLR).

Ac.Model	Total	F0	F1	F2	F3	F4	F5	FX
AM-6+auto cluster	29.8	18.8	27.0	39.1	29.9	36.3	30.1	54.2
AM-6+auto cluster +MLLR	27.8	17.9	25.8	33.1	26.6	35.2	27.8	49.9
AM-6+auto cluster + iterative MLLR	27.0	17.3	24.9	32.5	26.5	35.7	26.1	47.5

Table 13. Clustering for Unsupervised Adaptation (% WER): AM-6-auto - baseline with clustering, AM-6-auto cluster +MLLR1 - additionally one iteration of MLLR, M-6-auto cluster +iterative MLLR - iterative MLLR.

[8] R. Bakis et al., Transcription of BN Shows with the IBM LVCSR System", Proc. DARPA Sp. Reco. Workshop, 1997.

[9] F. Kubala et al., "The 1996 BBN Byblos Hub4 Transcription System", Proc. DARPA Sp. Reco. Workshop, 1997

[10] P. Placeway et al., "The 1996 Hub4 Sphinx System", Proc. DARPA Sp. Reco. Workshop, 1997.

[11] P. C. Woodland et al., "The Development of the 1996 HTK Broadcast News Transcription System", Proc. DARPA Sp. Reco. Workshop, 1997

[12] J. L. Gauvain et al., "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System", Proc. DARPA Speech Recognition Workshop, 1997.

[13] R. A. Gopinath, "Maximum Likelihood Modeling With Gaussian Distributions for Classification", Proceedings of ICASSP 1998.

[14] M. J. F. Gales, "Semi-tied Covariance Matrices", Proceedings of ICASSP 1998.

[15] S. Chen et al, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition", submitted to ICASSP 1998.

[16] C. J. Legetter et al., "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous density HMM's", Computer Speech and Language, vol. 9, no. 2, pp 171-186.

[17] J. L. Gauvain et al., "Maximum-a-Posteriori estimation for multivariate Gaussian observations of Markov chains", IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, Apr 1994.

[18] N. Kumar. "Investigation of Silicon-Auditory Models and Generalization of LDA for Improved Speech Recognition", PhD Thesis, Johns Hopkins Univ., 1997.

[19] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition", Tech. Rep., CUED/FINFENG/TR291, Cambridge Univ., 1997.

[20] M. J. F. Gales, "Semi-tied Full-covariance matrices for hidden Markov Models", Tech. Rep., CUED/FINFENG/TR287, Cambridge Univ., 1997.

[21] S. Chen and P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", Proceedings of the Speech Recognition Workshop, 1998.

[22] R. A. Gopinath, "Constrained Maximum Likelihood Modeling With Gaussian Distributions", Proceedings of the DARPA Speech Recognition Workshop, Lansdowne, VA, Feb 1998.